动态热门话题的“特征词条本体”自动构建与进化研究*

马 静 何雪枫 简旭文
(南京航空航天大学经济与管理学院 南京 210016)

摘要: 【目的】设计一种“特征词条本体”的自动构建及进化算法。【应用背景】热门话题产生的时间和话题演化往往是快速的, 且涉及领域广泛, 而现有的本体自动构建研究局限于具体领域的知识表达, 无法有效地对这种动态热门话题进行本体语义支持, 也不能进行有效跟踪与优化。【方法】通过对热门话题中关键事件的内容分析并由特征词组合而成的“特征词条本体”来描述热门话题的方法, 设计一种快速自动生成“特征词条本体”的算法; 在初始本体指导下, 利用话题跟踪结果进行“特征词条本体”进化算法的设计, 以满足不断更新的话题语义表述需求。【结果】针对热门话题“魏则西百度推广事件”, 使用爬虫工具采集 11 174 条新浪微博作为语料库进行实验, 抽取生成拥有 7 421 个特征词条、39 个特征词节点、781 个特征词关系的初始本体, 基于话题跟踪结果进化为拥有 24 564 个特征词条, 67 个特征词节点, 1 818 个特征词关系的进化本体, 其漏报率、误报率、损耗代价分别为 0.1261, 0.0964, 0.5985, 优于 TF-IDF 算法。【结论】“特征词条本体”的表述方式明显比单个词汇的本体表述准确率高, 且语义相似度更容易计算, 比较符合动态热门话题的快速语义处理。

关键词: 特征词条 本体生成 本体进化 话题跟踪
分类号: TP391 G353

1 引 言

随着互联网信息的爆炸, 海量文本的语义识别与表达成为主要难点。本体作为表达现实世界知识的重要方法得到研究者的极大关注[1]。本体被引入计算机科学中作为知识表示的方法并被广泛使用, 包括: 知识工程、智能信息处理、软件工程、自然语言处理等诸多领域, 并将成为语义网、基于知识的下一代智能计算、信息抽取和智能检索等许多领域的基础和关键[2]。由于本体构建中概念及其关系的建模大多都需要手工构建, 人工构建本体存在成本高、构建时间长并极其依赖专家的参与程度等一系列的问题, 成为本体构建的障碍[3]。

自动或半自动本体构建成为近年研究探索的热点。目前, 本体自动构建方式主要有三种:

(1) 通过聚类算法获取研究领域内的概念与关系进行构建。Lin 等[4]在其本体自动构建研究中通过 CBC(Clustering By Committee)聚类发现领域概念; Srivastava 等[5]研究了从文本信息中获取本体的层次以及关联关系, 分别使用相似度度量聚类(Similarity-based Clustering)、集合理论聚类(Set-theoretic Clustering)两种方式进行本体关联的挖掘聚类研究, 并分析了这两种方法的聚类有效性、效率和可跟踪性; 何婷婷等[6]提出了一种多重聚类技术自动构造本体的方法。

(2) 根据已有的词典或术语表自动构建本体。He 等[7]与 Lim 等[8]给出了获得概念分层语义关系的方法:

通讯作者: 马静, ORCID: 0000-0001-8472-2518, E-mail: majing5525@126.com。

*本文系国家自然科学基金面上项目“基于演化本体的网络舆情自适应话题跟踪方法研究”(项目编号: 71373123)、江苏高校哲学社会科学研究重点项目“基于超网络的江苏教育微博舆情多元意见演化模型及应用研究”(项目编号: 2015ZDIXM007)、高校重大项目培育基金“基于‘模型–数据双驱动’的复杂社会网络行为大数据分析方法研究”(项目编号: NP201630X)的研究成果之一。

从语料库中抽取术语，分别根据已有术语表或词典计算术语之间的相似性和语义分类层次，同时结合词汇在表达相同主题的各文本中所具有的重要性，构建各术语之间的语义关系；马静等[9]选择使用 NASA 叙词表，将广泛的航空产品类概念进行本体映射，构建航空产品的领域本体；唐爱民等[10]提出利用半结构化文本作为本体的知识源，基于词汇功能语法理论对句子进行分析，将原文中语法表述的文字转换成语义表述，从而获取本体。

(3) 通过网络或者领域图来挖掘概念间的层次关系进行本体构建。Chen 等[11]提出基于自适应谐振网络和贝叶斯网络的领域本体自动构建算法；侯鑫等[12]提出一种基于图上随机游走的词汇加权算法，获取候选概念进而自动构建领域本体。郑学伟[13]采用基于图的构建原理，在关系运算中采用基于频繁信息子图的 gSpan 算法得到本体。

上述研究中，第一种研究方向可以将相似概念聚集在一起，但是无法获得概念与概念之间的关系描述；第二种研究方向利用语言分析技术涉及的学科多且跨度大，完全实现本体自动构建很困难；第三种研究方向随着领域图或网络中顶点数量的增加，本体概念与关系抽取的准确率就会下降，自动构建本体尚不能满足实际应用的需求。同时，随着网络舆情事件的影响力越来越大，面向动态热门话题的自动本体构建需求迫切，而现有的本体构建方法都因局限于某些已知领域的积累，因而无法快速产生热门话题的相关本体。

本文提出一种“特征词条本体”的概念，从首发新闻报道中快速自动生成一个热门话题的核心语义，实现对话题的语义表述；基于词汇间共现关系设计“特征词条本体”抽取生成算法；在此基础上设计基于不断演化的话题跟踪结果下“特征词条本体”进化算法。

2 “特征词条本体”自动生成的算法设计

2.1 “特征词条本体”的提出

对于动态热门话题中的新闻内容，有些词语与目标话题关系密切，有些词语则和话题没有太大关系，单单计算词语的词频无法获知新闻与话题关系的密切程度。然而新闻内容中某些具有特定关系的词组却能在很大程度上体现一个话题的语义特性，例如在“魏则西百度推广事件”的话题中，(魏则西，百度，事件)，(百度，推广，事件)等词语组合几乎会在所有新闻内容中出现，这类词组表示该话题的关键事件进而表述话题的核心语义，同时若一条待检测的新闻中频繁包含该类词组，其与目标主题相关的可能性就很大。其次通过对不同概率分布的特征词条的区分，还能概括出该话题下不同子事件，例如在魏则西百度推广事件的话题中(魏则西，莆田，医院)，(医院，莆田，责任)等词条在某些新闻内容中出现频率较高，反映了目标话题下不同侧重点的子事件。

本文将能够表示话题中关键事件语义的特征词组合定义为特征词条，其数学符号为 c, $c=\{w_1, w_2\cdots\}$，其中 $w_1, w_2\cdots$为组成该词条的特征词。由特征词条组成并用来描述话题的集合，本文将其定义为“特征词条本体”，其数学符号为 C, $C=\{c_1, c_2\cdots\}$。

2.2 特征词条抽取与初始本体生成

特征词条是利用特征词之间的共现关系，将共现概率高的几个特征词组合在一起以表示话题中的特定内容。本文利用特征词之间的互信息值来计算词的共现概率，词 w_i 与 w_j 的共现概率 $M(w_i, w_j)$ 计算公式如下：

$$M(w_i, w_j) = \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \quad (1)$$

其中，$p(w_i, w_j)$ 为特征词 w_i 与 w_j 在同一句子中出现的频率，$p(w_i)$ 与 $p(w_j)$ 分别为 w_i 与 w_j 在训练集中各自的出现频率。

“特征词条本体”抽取生成的过程如下：

(1) 创建临时词组集合 G_1，G_2，G_3，两两计算特征词互信息值 $M(w_i, w_j)$，当值大于阈值 T_m 时，将词组 (w_i, w_j) 加入词组集合 G_1 中。

(2) 对临时集合 G_1 进行判断，若 G_1 为空，结束抽取过程，此时“特征词条本体” C 为特征词的集合，若 G_1 不为空，则继续特征词条的抽取。

(3) 取 G_1 中第一个特征词组 (w_i, w_j)，遍历 $w_{i+1}, w_{i+2}\cdots w_n$，若发现存在 $M(w_i, w_k)>T_m$，$M(w_j, w_k)>T_m$，则将 (w_i, w_j, w_k) 加入 G_2，否则将 (w_i, w_j) 加入到集合 G_3，并将 (w_i, w_j) 从 G_1 中去除。

(4) 重复步骤(3)，直至临时集合 G_1 为空，并对 G_2 集合进行判断，若 G_2 为空，结束抽取，集合 G_3 即

为“特征词条本体”C，若 G_2 不为空，则继续特征词条抽取。

(5) 将 G_2 集合中特征词条加入到 G_1，将 G_2 清空，重复以上步骤寻找特征词条，直至临时集合 G_1 为空时，G_2 也为空，此时集合 G_3 即为“特征词条本体”C。

2.3　特征词节点权重的生成

得到“特征词条本体”C 后，需要计算单个特征词在“特征词条本体”中的权重，以表述该特征词的重要程度。本文从网络图的角度出发，将“特征词条本体”集合中的词条看成是由特征词组成的网络图，利用词与词之间的关系(即网络图中的两个特征词节点间的连线)计算单个特征词权重。

首先根据词条出现的次数计算词条权重，则词条 c_i 在训练集中的权重计算公式为:

$$Q(c_i)=\frac{t_i}{\sum t_i} \tag{2}$$

其中，t_i 是词条 c_i 在训练集中出现的次数，$\sum t_i$ 是所有词条在训练集中出现的次数之和，两者相除即为词条 c_i 的权重。在此基础上，计算两个特征词 w_i,w_j 之间关系(即网络图中两点的连边)的权重，其计算公式如下:

$$Q(edge_{ij})=\frac{\sum\{Q(c_i),Q(c_j)\cdots\}}{\sum Q(edge_{ij})} \tag{3}$$

其中，$\{Qc_i,Qc_j\cdots\}$ 是词条本体中出现了 w_i,w_j 连接关系的特征词条集合，$\sum\{Q(c_i),Q(c_j)\cdots\}$ 是集合 $\{Qc_i,Qc_j\cdots\}$ 中词条权重之和，并进行权重的归一化处理，除以网络图中所有关系的权重和 $\sum Q(edge_{ij})$，得到 w_i,w_j 关系权重 $Q(edge_{ij})$。利用公式(3)，可以求出网络图中有连线的两两特征词关系的权重，在此基础上，求出网络图中每个特征词节点的权重。特征词 w_i 的权重计算公式如下:

$$Q(w_i)=\frac{\sum\{Q(edge_{ij}),Q(edge_{ik})\cdots\}}{k} \tag{4}$$

其中，$\{Q_{edge_{ij}},Q_{edge_{ik}}\cdots\}$ 是网络图中与特征词节点 w_i 关联边的集合，$\sum\{Q(edge_{ij}),Q(edge_{ik})\cdots\}$ 是集合 $\{Q_{edge_{ij}},Q_{edge_{ik}}\cdots\}$ 中关联边的权重之和，并除以特征词节点 w_i 的度 k，即求出“特征词条本体”中特征词 w_i 的权重 $Q(w_i)$。

3　基于话题跟踪的“特征词条本体”进化的算法设计

本体构建后需要不断添加新的概念以满足实际需求，因此需要依据相应的理论、方法及标准对本体的概念、结构及关系进行完善，即本体进化[14-15] (Ontology Evolution)。“特征词条本体”同样需要在话题跟踪结果的基础上不断添加新的语料，实现其自身的概念与关系的完善。

本体进化的基础是加入大量话题相关的语料，本文采用话题跟踪的方式寻找相关新闻:

(1) 计算待检测新闻 d 与现有“特征词条本体”C 的相似度值，使用向量空间模型(Vector Space Model, VSM)分别描述新闻 d 与“特征词条本体”C，将 d 与 C 的相似度计算抽象成两个对应向量的相似度计算，并用数学上的向量余弦公式定量化计算为:

$$sim(d,C)=\frac{\sum_{i=1}^{n}Q_d(w_i)\times Q(w_i)}{\sqrt{(\sum_{i=1}^{n}Q_d^2(w_i))(\sum_{i=1}^{n}Q^2(w_i))}} \tag{5}$$

其中，$Q_d(w_i)$ 为 w_i 在待检测新闻 d 中出现的频率，$Q(w_i)$ 为公式(4)计算的特征词 w_i 的权重，n 为向量空间中出现特征词的数量。

(2) 将相似度值计算结果 sim(d,C) 与设定判断阈值 T_d 相比较，判断它是否为主题相关，若结果大于判断阈值 T_d，则判定新闻 d 是目标话题 D 的相关内容。

本文在算法中嵌入一个比判断阈值 T_d 大的进化阈值 T_u，如果相似度 sim(d,C) 大于 T_u，则认为该新闻 d 不仅话题相关，而且可以描述话题 D，将其加入“特征词条本体”的训练集中。完成话题检测与跟踪之后，利用新的训练集快速抽取生成进化本体，基于话题跟踪的“特征词条本体”进化算法思路如图 1 所示:

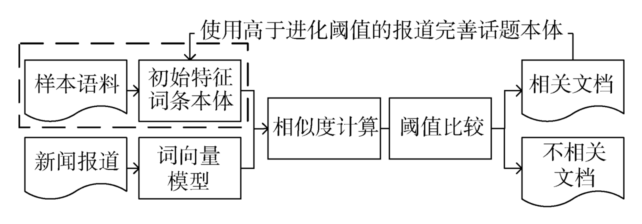


图 1　基于话题跟踪结果的“特征词条本体”进化思路

4 实验设计及分析

依据算法进行实验，实验环境使用的是南京航空航天大学信息管理与电子商务研究所的数据挖掘与语义分析研究平台。根据提出的整体研究思路，选择近期热门的“魏则西百度推广事件”作为目标话题，以新浪微博内容作为新闻语料，进行“特征词条本体”的自动生成与进化实验。在实验结果分析中，为了验证“特征词条本体”在话题语义表述上的有效性，选取基于词频的 TF-IDF 算法进行话题跟踪与检测的对比实验，使用 TDT4[16]作为标准，评测两种方法的跟踪效果。

4.1 实验数据

(1) 微博数据的采集

实验需要两种类型的微博：与“魏则西百度推广事件”相关的和不相关的微博。为了采集实验所需的微博数据，使用研究平台中的新浪微博抓取爬虫于 2016 年 5 月 13 日至 2016 年 5 月 22 日连续 10 天对微博语料进行增量爬取，去除字数过少微博与重复微博(指微博 id 重复，而不是内容重复)，共计获得微博 11 174 条。其中 5 月 13 日至 5 月 17 日以微博系统自身的话题划分为依据爬取“魏则西百度推广事件”相关微博共计 4 480 条，作为初始词条本体生成的训练集；于 5 月 18 日至 5 月 22 日爬取热门微博共计 6 694 条，作为实验的测试集，用来验证“特征词条本体”在语义表述上的准确性，并在此基础上进行词条本体的进化。

为了判断测试集中的热门微博是否与“魏则西百度推广事件”话题相关，采用人工观察的方式，由两个观察员共同评定测试集中热门微博的相关性。本次实验中微博采集的时间分布及是否与话题相关的结果如表 1 所示：

表 1 微博的时间分布与话题相关性结果

对比项	训练集数据					测试集数据				
	13 日	14 日	15 日	16 日	17 日	18 日	19 日	20 日	21 日	22 日
主题相关的微博数量	881	905	947	888	859	458	459	493	524	652
主题不相关的微博数量	0	0	0	0	0	780	751	781	949	847
总计	881	905	947	888	859	1 238	1 210	1 274	1 473	1 499

(2) 微博语料的预处理

本次实验预处理分三个步骤依序进行：使用 NLPIR-ICTCLAS2016 系统①(Institute of Computing Technology, Chinese Lexical Analysis System)进行微博内容的分词，导入“百度汉语分词词库”，提高了分词准确性；分词之后会有大量的语气词、助词，比如“哪里”、“其他”、“是的”等词汇没有任何实际意义，使用“哈工大停用词表”、“四川大学机器智能实验室停用词库”、“百度停用词列表”等综合去除语料库的停用词；对出现的一些网站引用、乱码等信息，比如“@”、“http”、“.com”、“#”等，使用正则表达式进行匹配去除无关词汇。经过文本预处理，实现了对微博语料库的分词与去杂，得到词汇 21 634 个。

4.2 实验方法

本文使用训练集的数据实现初始本体的自动生成，使用初始本体在测试集中寻找高于进化阈值 T_u 的微博，利用跟踪结果实现词条本体的进化。

(1) 初始特征词条本体的生成

在训练集中去除出现次数少、对内容表达没有太大意义的低频词，使用微博出现频率大于阈值 T_w 的名词或动词作为特征词，此处阈值 T_w 取 0.01，然后按照特征词条抽取算法，抽取共现概率高于阈值 T_m 的特征词组成特征词条进而生成“特征词条本体”C，此处阈值 T_m 取 0.015。C 中部分特征词条如图 2 所示：

[事件/n,推广/vn,魏则西/nr] [推广/vn,是/vshi,百度/nz]
[事件/n,推广/vn,百度/nz] [推广/vn,是/vshi,魏则西/nr]
[推广/vn,百度/nz,魏则西/nr] [事件/n,推广/vn,是/vshi]
[事件/n,是/vshi,百度/nz] [事件/n,看/v,魏则西/nr]
[事件/n,是/vshi,魏则西/nr] [事件/n,百度/nz,看/v]
[是/vshi,百度/nz,魏则西/nr] [百度/nz,看/v,魏则西/nr]

图 2 初始“特征词条本体”中部分特征词条结果

在初始“特征词条本体”的基础上，依据特征词节点权重的生成算法，依次生成词条权重、关系权重、与特征词权重。并对比传统基于词频的 TF-IDF 算法

①https://github.com/NLPIR-team/NLPIR/tree/master /NLPIR-ICTCLAS.

的特征词权重，对比结果如图 3 所示：

事件/n	0.085400
推广/vn	0.070311
百度/nz	0.069240
看/v	0.061486
是/vshi	0.055976
魏则西/nr	0.047372
来/vf	0.047012
医院/n	0.038451
这/rzv	0.037203
度/qv	0.033184
能/v	0.029045
人/n	0.028235
事/n	0.022512
全文/n	0.022232
有/vyou	0.021270

(a) TF-IDF算法

百度/nz	0.108776
推广/vn	0.096503
是/vshi	0.080982
事件/n	0.078995
魏则西/nr	0.073033
有/vyou	0.045181
人/n	0.036250
看/v	0.033398
能/v	0.030082
医院/n	0.029677
来/vf	0.027885
这/rzv	0.023919
事/n	0.021894
没有/v	0.019013
去/vf	0.018795

(b) 本体方法

图 3　权重最高的 20 个特征词权重对比图

通过对比两种方法，可以看出，在使用本文方法后，“魏则西百度推广事件”话题中“百度”、“魏则西”等特征词的权重提高，而与话题相关度较低的特征词如“看”、“度”、“全文”的权重明显降低。使用 Gephi①软件对词条本体中的特征词关系进行可视化展示，依序处理特征词条，如[事件/n，百度/nz，魏则西/nr]，按照特征词条的顺序，以箭头进行串联，生成该词条对应的线路：事件/n->百度/nz ->魏则西/nr，处理完全部的特征词条，即可生成初始词条本体的特征词关系如图 4 所示：

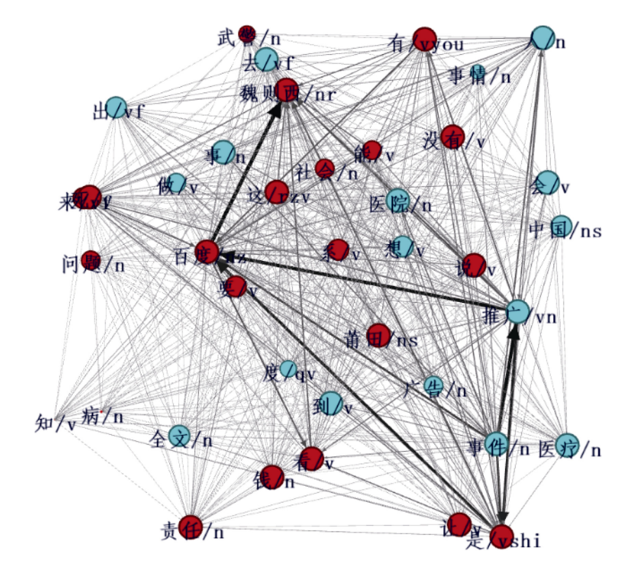


图 4　初始词条本体的特征词关系图

(2) 基于跟踪结果的本体进化

在初始“特征词条本体”指导下对测试集进行话题跟踪，使用公式(5)计算一条微博与“特征词条本体”的相似度。因为微博本身的内容较短并且需要选择相似度高的微博对词条本体进行进化，如果微博数量太少的话会影响后续本体进化的效果，所以结合实验结果，设置判断阈值 T_d =0.3，进化阈值 T_u =0.4，可得词条本体在测试集中的跟踪结果如表 2 所示：

表 2　测试集中微博的时间分布与话题跟踪结果

对比项	18 日	19 日	20 日	21 日	22 日
相似度大于判断阈值的微博数量	468	450	522	529	647
相似度大于进化阈值的微博数量	132	130	122	104	200

选择 18 日至 22 日中相似度大于进化阈值 T_u 的 688 条微博作为话题跟踪结果对“魏则西百度推广事件”的话题本体进行改进与完善，实现本体进化，进化后的特征词关系如图 5 所示：

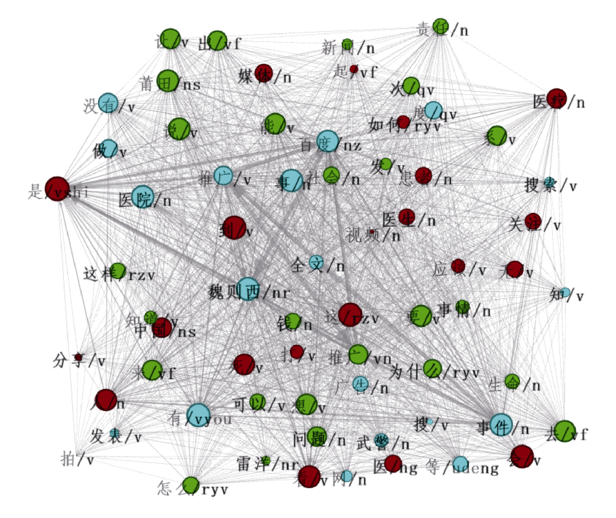


图 5　本体进化后的特征词关系图

对比图 4 与图 5，可以看出，进化后的“特征词条本体”在表述新闻话题时语义更加丰富：特征词与特征词关系的数量明显增长，以本次“魏则西百度推广事件”为例，初始“特征词条本体”共有特征词条 7 421 个，特征词节点 39 个，特征词关系 781 个，在经过本体进化后，共有特征词条 24 564 个，特征词节点 67 个，特征词关系 1 818 个；更加能够反映话题下子事件的

①https://gephi.org/.

语义，例如观察进化后的特征词关系图中[事件/n，发表/v，魏则西/nr]、[发表/v，百度/nz，魏则西/nr]、[中国/ns，发表/v，莆田/ns]、[媒体/n，发表/v，责任/n]、[媒体/n，发表/v，责任/n]等词条可以获知该话题下关于媒体问责魏则西事件的相关内容。

4.3 实验分析

(1) 话题跟踪实验的设计

为了验证“特征词条本体”在语义表达上的有效性，本文设计了与 TF-IDF 算法的对比实验，比较两种方法在话题跟踪上的效果，效果更好的方法即为在语义表述上更好的方法。

基于 TF-IDF 算法的话题跟踪与本体方法指导下的话题跟踪流程大致相同，都依赖于公式(5)的相似度计算及阈值判定，唯一不同的是特征词权重的确定，TF-IDF 是在词频的基础上，对每个词分配一个“重要性”权重，其计算公式如下：

$$Q_{tf\text{-}idf}(w_i)=\frac{t_{wi}}{\max(t)}\times\log\left(\frac{allDoc}{doc(w_i)+1}\right) \quad (6)$$

其中，t_{wi} 为特征词 w_i 在语料中出现的次数，max(t) 为语料中出现次数最多的词的次数，两者相除即为词频(Term Frequency)，allDoc 是语料中文档总数，$doc(w_i)$ 是包含 w_i 的文档总数，求对数之后即为逆文档频率(Inverse Document Frequency)，词频与逆文档频率的乘积即为 w_i 的 TF-IDF 权重。

(2) 话题跟踪判断标准

NIST 为 TDT 建立了一套完整的评测体系[17]，使用损耗代价 $(C_{Det})_{Norm}$ 作为系统的评价指标，此值越小则表示系统性能越好，其计算公式如下：

$$C_{Det}=C_{Miss}\times P_{Miss}\times P_{Target}+C_{FA}\times P_{FA}\times P_{non\text{-}Target} \quad (7)$$

$$(C_{Det})_{Norm}=\frac{C_{Det}}{\min(C_{Miss}\times P_{Target},C_{FA}\times P_{non\text{-}Target})} \quad (8)$$

其中，C_{Miss} 和 C_{FA} 分别代表漏报率和误报率的代价系数，在 TDT4 中 C_{Miss} 和 C_{FA} 分别取值为 1 和 0.1，认为漏报代价比误报代价高很多；P_{Target} 和 $P_{non\text{-}Target}$ 是先验目标概率($P_{non\text{-}Target}=1-P_{Target}$)，$P_{Target}$ 一般取 0.02[18]，P_{Miss} 和 P_{FA} 分别是系统漏报率和误报率，两者均是越小越好，但是两者之间存在一定的矛盾，一般情况下漏报率较低的误报率会较高，而误报率较低的漏报率会较高，计算公式如下：

$$漏报率P_{Miss}=\frac{系统中未识别的相关微博数}{语料库中描述该话题的微博总数}\times 100\% \quad (9)$$

$$误报率P_{FA}=\frac{判定为相关的不相关微博数}{语料库中与该话题不相关的微博总数}\times 100\% \quad (10)$$

(3) 话题跟踪结果分析

使用判断阈值($T_d=0.3$)判断微博内容是否主题相关，对两种话题跟踪方法的第一次话题跟踪实验结果对比如表 3 所示：

表 3 第一次话题跟踪结果分析

对比项	18 日	19 日	20 日	21 日	22 日	总计
TF-IDF 未识别的相关微博数量	92	75	89	94	118	468
本文方法未识别的相关微博数量	79	53	46	87	139	404
TF-IDF 识别的不相关微博数量	51	71	63	93	136	414
本文方法识别的不相关微博数量	89	44	75	92	134	434
测试集中实有相关微博数量	458	459	493	524	652	2 586
测试集中实有不相关微博数量	780	751	781	949	847	4 108

按照 TDT4 评价方法，分别计算两种方法话题跟踪结果的漏报率、误报率和损耗代价并进行对比分析，结果如表 4 所示：

表 4 TF-IDF 与本体方法的话题跟踪结果的漏报率、误报率、损耗代价对比

指标	TF-IDF 方法	本体方法
漏报率 P_{Miss}	0.1810	0.1562
误报率 P_{FA}	0.1008	0.1056
损耗代价 $(C_{Det})_{Norm}$	0.6749	0.6736

从表 4 的结果可以看出：基于词条本体的话题跟踪方法比 TF-IDF 方法的漏报率有一定的下降，但是误报率与损耗代价并没有明显的提升，综合来看，初始词条本体的跟踪结果与 TF-IDF 方法效果相当。

然后将寻找到的相似度大于进化阈值的微博加入到词条本体的训练集，重新训练出进化的“特征词条本体”并进行第二次话题跟踪实验，实验结果如表 5 所示。

表 5 第二次话题跟踪结果分析

对比项	18 日	19 日	20 日	21 日	22 日	总计
本体方法未识别的相关微博数量	69	29	46	95	87	326
本体方法识别的不相关微博数量	83	54	73	82	104	396

计算进化后"特征词条本体"方法的话题跟踪结果的漏报率、误报率和损耗代价，并与表 4 中初始本体、TF-IDF 方法的计算结果进行对比分析，如表 6 所示：

表 6 三种方法的漏报率、误报率、损耗代价对比

指标	TF-IDF 方法	初始本体	进化本体
漏报率 P_{Miss}	0.1810	0.1562	0.1261
误报率 P_{FA}	0.1008	0.1056	0.0964
损耗代价 $(C_{Det})_{Norm}$	0.6749	0.6736	0.5985

从表 6 的实验结果可以看出：经过进化之后的词条本体在漏报率、误报率、损耗代价上都要优于前两种方法，表明进化后的词条本体的话题跟踪效果更优；其中漏报率的性能显著提高，这主要是由于进化后的"特征词条本体"的节点和关系大大增加，更能表示目标话题语义信息，能够更准确地跟踪话题；就损耗代价而言，根据损耗代价的计算公式，TDT 评测更加重视误报率对评测结果的影响，因此误报率较小的差别导致在损耗代价之间的差值没有两者漏报率之间的差值明显；本体方法对于话题跟踪结果的误报率并没有明显的提高，后续对算法的改进应集中在维持当前低漏报率水平的情况下，尽量减少误报率。

本次实验结果显示基于"特征词条本体"的话题跟踪的效果是优于 TF-IDF 算法，并且通过本体进化，能够进一步优化话题跟踪结果。因此，证明了本文提出的基于词关系的"特征词条本体"在动态热门话题的语义表述上是有效的。

5 结 语

本文从动态热门话题新闻中的词汇出发，将共现频率高的特征词组合成特征词条设计生成"特征词条本体"，在初始本体的指导下进行话题跟踪，利用跟踪结果对话题本体的概念和关系进行改进，完成"特征词条本体"的进化。通过实验证明了"特征词条本体"是一种表述更加准确的语义模型，并且可以满足动态热门话题快速表达的实际需求。但是目前本体语义表述的效果过于受人工阈值设置的影响，在未来的研究中可以考虑引入深度学习相关的模式，以尽可能降低人工干预进而提高"特征词条本体"的语义表达准确性。

参考文献：

[1] Studer R, Benjamins V R, Fensel D. Knowledge Engineering: Principles and Methods[J]. Data & Knowledge Engineering, 1998, 25(1): 161-197.

[2] 杜小勇，李曼，王珊．本体学习研究综述[J]．软件学报，2006, 17(9): 1837-1847. (Du Xiaoyong, Li Man, Wang Shan. A Survey on Ontology Learning Research[J]. Journal of Software, 2006, 17(9): 1837-1847.)

[3] 尚新丽．国外本体构建方法比较分析[J]．图书情报工作，2012, 56(4): 116-119. (Shang Xinli. Comparative Analysis of Foreign Ontology Construction Methods [J]. Library and Information Service, 2012, 56(4): 116-119.)

[4] Lin D, Pantel P. Induction of Semantic Classes from Natural Language Text[C]. In:Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA. 2001: 317-322.

[5] Srivastava S, Lamadrid J G. Extracting an Ontology from a Document Using Singular Value Decomposition [R]. Association of Computer and Information Science and Engineering Departments at Minority Institutions, 2001.

[6] 何婷婷，张小鹏．特定领域本体自动构造方法[J]．计算机工程，2007, 33(22): 235-237. (He Tingting, Zhang Xiaopeng. Approach to Automatical Construction of Domain Ontology [J]. Computer Engineering, 2007, 33(22): 235-237.)

[7] He T T, Zhang X P, Ye X H.An Approach to Automatically Constructing Domain Ontology[C]. In: Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation, Wuhan, China.2006:150-157.

[8] Lim S Y, Park S B, Lee S J. Constructing an Ontology Based on Terminology Processing [C]. In: Proceedings of the 9th International Conference on Knowledge-Based Intelligent Information and Engineering Systems. Springer, 2005: 304-310.

[9] 马静，吴一占，刘思峰．基于领域本体的信息抽取模式生成与系统实现[J]．情报学报，2008, 27(2): 193-198. (Ma Jing, Wu Yizhan, Liu Sifeng. Domain Ontology-based Information Extraction [J]. Journal of the China Society for Scientific and Technical Information, 2008, 27(2): 193-198.)

[10] 唐爱民，真溱，樊静．基于叙词表的领域本体构建研究[J]．现代图书情报技术，2005 (4): 1-5. (Tang Aimin, Zhen Zhen,

Fan Jing. Thesaurus-based Approach to Build Domain Ontology[J]. New Technology of Library and Information Service, 2005 (4): 1-5.)

[11] Chen R C, Chuang C H. Automating Construction of a Domain Ontology Using a Projective Adaptive Resonance Theory Neural Network and Bayesian Network [J]. Expert Systems, 2008, 25(4): 414-430.

[12] 侯鑫, 张旭堂, 金天国, 等. 面向知识与信息管理的领域本体自动构建算法[J]. 计算机集成制造系统, 2011, 17(1): 159-170. (Hou Xin, Zhang Xutang, Jin Tianguo, et al. Automatic Construction of Domain Ontology Oriented to Knowledge and Information Management [J]. Computer Integrated Manufacturing Systems, 2011, 17(1): 159-170.)

[13] 郑学伟. 基于知识管理的本体自动构建算法研究[J]. 计算机技术与发展, 2014, 24(12): 64-69. (Zheng Xuewei. Research on Ontology Automatic Construction Algorithm Based on Knowledge Management [J]. Computer Technology and Development, 2014, 24(12): 64-69.)

[14] 马文峰, 杜小勇. 领域本体进化研究[J]. 图书情报工作, 2006, 50(6): 71-75. (Ma Wenfeng, Du Xiaoyong. A Study on Domain Ontology Evolution[J]. Library and Information Service, 2006, 50(6): 71-75.)

[15] 杜小勇, 马文峰, 武文娟. 学科领域本体的构建与进化——以经济学领域本体为例[J]. 现代图书情报技术, 2007(3): 7-12. (Du Xiaoyong, Ma Wenfeng, Wu Wenjuan. Construction and Evolution of Discipline Domain Ontology——A Case Study for Economics Domain Ontology[J]. New Technology of Library and Information Service, 2007(3): 7-12.)

[16] 洪宇, 张宇, 刘挺, 等. 话题检测与跟踪的评测及研究综述[J]. 中文信息学报, 2007, 21(6): 71-96. (Hong Yu, Zhang Yu, Liu Ting, et al. Topic Detection and Tracking Review[J]. Journal of Chinese Information Processing, 2007, 21(6): 71-96.)

[17] 焦健, 瞿有利. 知网的话题更新与跟踪算法研究[J]. 北京交通大学学报, 2009, 33(5):132-136. (Jiao Jian, Qu Youli. Algorithm Study of Topic Tracking Based on HowNet and Topic Renewal [J]. Journal of Beijing Jiaotong University, 2009, 33(5):132-136.)

[18] 洪宇, 仓玉, 姚建民, 等. 话题跟踪中静态和动态话题模型的核捕捉衰减[J]. 软件学报, 2012, 23(5): 1101-1119. (Hong Yu, Cang Yu, Yao Jianmin, et al. Descending Kernel Track of Static and Dynamic Topic Models in Topic Tracking [J]. Journal of Software, 2012, 23(5): 1101-1119.)

作者贡献声明:

马静: 提出研究思路和方案, 论文审阅与修订;
何雪枫: 扩展研究思路, 进行实验, 论文撰写与修订;
简旭文: 扩展研究思路, 辅助进行实验。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据见期刊网络版 http://www.infotech.ac.cn。

[1] 马静, 何雪枫, 简旭文. 实验所用到的微博数据 1.xls. 通过爬虫爬取的数据.

[2] 马静, 何雪枫, 简旭文. 实验所用到的微博数据 2.xls. 通过中文预处理的数据.

[3] 马静, 何雪枫, 简旭文. 自动抽取生成的特征词条.xls. 初始本体.

[4] 马静, 何雪枫, 简旭文. 自动抽取生成的特征关系.xls. 初始本体.

[5] 马静, 何雪枫, 简旭文. 自动抽取生成的特征词权重.xls. 初始本体.

[6] 马静, 何雪枫, 简旭文. 自动抽取生成的特征词条.xls. 进化本体.

[7] 马静, 何雪枫, 简旭文. 自动抽取生成的特征关系.xls. 进化本体.

[8] 马静, 何雪枫, 简旭文. 自动抽取生成的特征词权重.xls. 进化本体.

[9] 马静, 何雪枫, 简旭文. 实验的数据库文件.sql. 数据库.

[10] 马静, 何雪枫, 简旭文. 程序源码.zip. 源码.

收稿日期: 2016-06-12
收修改稿日期: 2016-07-25

Automatically Building "Feature Items Ontology" for Trending Topics

Ma Jing　He Xuefeng　Jian Xuwen
(College of Economic and Management, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)

Abstract: [**Objective**] This paper aims to propose an algorithm to build "Feature Items Ontology". **[Context]** Trending topics online are constantly changing and involve extensive fields. The existing research on automatically creating Ontology is limited to specific areas, which cannot effectively process the dynamic trending topics. **[Methods]** First, we analyzed the contents of major events from the trending topics. Second, we designed an algorithm automatically generating the Ontology. Third, with the guidance of initial Ontology, proposed an evolutionary algorithm to track the changing topics. **[Results]** Using the case of "Wei Zexi and Baidu" as an example, we collected 11,174 Sina Weibo posts to conduct two rounds of experiment. We initially extracted 7,421 feature items, 39 key nodes, and 781 key relationships. For the evolutionary results, we got 24,564 feature items, 67 key nodes, and 1,818 key relations. The missing rates, the false positive rates, and the loss costs were 0.1261, 0.0964 and 0.5985, which were all better than those of the TF-IDF algorithm. **[Conclusions]** The "Feature Items Ontology" is more accurate than the single word Ontology description, and is easier to calculate the semantic similarity. It is an appropriate method to retrieve semantic information from the dynamic trending topics.

Keywords: Feature items　Ontology generation　Ontology evolution　Topic tracking

Mellon 基金会资助 BitCurator 进行扩展，以改善对数字原生资源的分析和访问功能

北卡罗来纳大学教堂山分校于近日从 Andrew W. Mellon 基金会获得一笔 75 万美元的基金资助，用于对 BitCurator NLP 进行扩展。BitCurator NLP 项目旨在开发能将自然语言处理(NLP)方法应用于数字原生图书馆馆藏、档案馆馆藏和博物馆馆藏的相关软件和协议。项目为期两年，所创建的新工具将使得图书馆、档案馆和博物馆领域的专业人员能够更有效和更高效地流通数字馆藏资源，并最终使得用户在搜索信息或文档时更容易发现并访问这些馆藏资源。

BitCurator NLP 将以 BitCurator 和 BitCurator Access 项目为基础，这两个项目旨在开发并分发工具以帮助图书馆、档案馆和博物馆管理快速增长的具有文化价值的数字资源。

BitCurator 开发了一个开源软件环境，便于将资源从便携式媒体(如软盘、闪存驱动器和硬盘驱动器)迁移到更可持续的环境。用户可以创建磁盘映像，分析文件和文件系统，提取数据和元数据，以及识别和编辑敏感信息，等等。

BitCurator Access 通过 BCA Webtools 进一步增强了 BitCurator 的功能，允许用户动态浏览磁盘映像的文件系统，以及搜索许多常见文件类型的内容。BitCurator Access 还开发了用于修改敏感信息的工具，并尝试使用仿真作为磁盘映像内容的访问机制。BitCurator 和 BitCurator Access 的产品和相关社区由独立的、成员驱动的 BitCurator 联盟在维护。

BitCurator NLP 将生成一个开源软件，用于提取、分析和生成馆藏中数字资源文本的相关特征的报告。该软件还将帮助图书馆、档案馆和博物馆改进或实施 NLP 功能，以从数字馆藏中读取文件，并为最终用户按需生成报告。

(编译自: https://librarytechnology.org/news/pr.pl?id=21961)

(本刊讯)